

www.discover-earth.org

Service Oriented Architecture for Harvesting Distributed Data Repositories

Helen Conover, Sara Graves, Ken Keiser, Lamar Hawkins
Information Technology & Systems Center
University of Alabama in Huntsville
Huntsville, AL 35899

www.itsc.uah.edu

Presented by Danny Hardin
NASA Earth Science Technology Conference
College Park, MD
27-29 June 2006

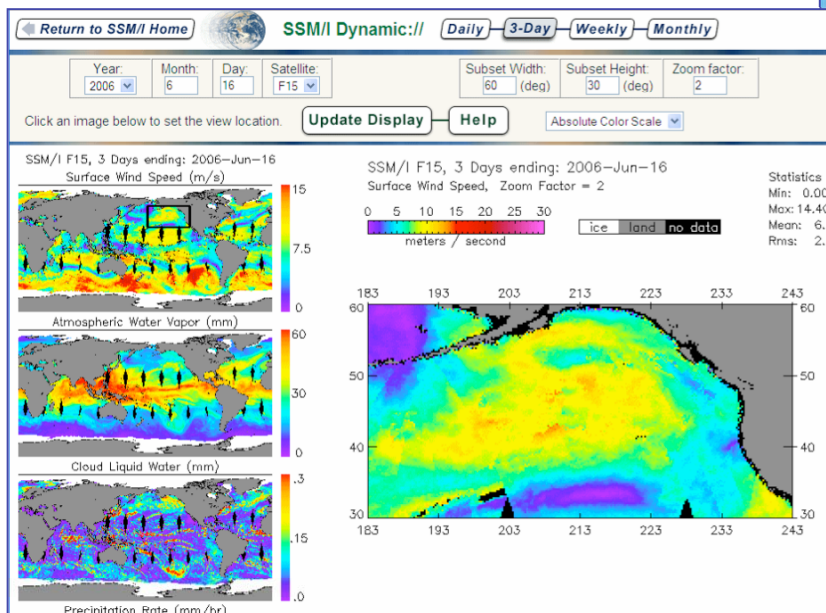


Remote Sensing Systems
www.remss.com



DISCOVER

The primary objective of the DISCOVER Project is to provide highly accurate, long-term ocean and climate products suitable for the most demanding Earth research applications via easy-to-use display and data access tools.



DISCOVER

Carefully Calibrated, Long-Term Ocean and Climate Data Records

Oceanic & Atmospheric Data

www.discover-earth.org

DISCOVER is a NASA Earth Science REASoN project.

REASoN (Research, Education and Applications Solution Network) is a distributed network of data and information providers for NASA's Earth Science Enterprise (ESE) Science, Applications and Education programs.

DISCOVER core mission: provide highly accurate, long-term climate data records and near-real-time ocean products suitable for the most demanding Earth research applications via easy-to-use display and data access tools.

[Home](#)

[Project Info](#)

[Data Services](#)

[Information Technology](#)

[DISCOVER IT Development at UAH](#)

[Top](#)

DISCOVER

Carefully Calibrated, Long-Term Ocean and Climate Data Records

The DISCOVER Project

The primary objective of the Distributed Information Services for Climate and Ocean Products and Visualizations for Earth Research (DISCOVER) Project is to provide highly accurate, long-term ocean and climate products suitable for the most demanding Earth research applications via easy-to-use display and data access tools. These products are derived from a large network of satellite microwave sensors going back to 1979. Most of the products are produced in near real-time (3-12 hours) on a 24/7 basis and hence are also suitable for some weather applications. The products include sea surface temperature and wind, air temperature, atmospheric water vapor, cloud water, and rain rate. A key element of DISCOVER is the merging of multiple sensors from multiple platforms into geophysical data sets consistent in both space and time.

The IT goals for DISCOVER include:

- On-line services for data access and visualization
- Interoperability technologies for improved usability
- Flexible architecture to adapt to changing user requirements
- IT Approach
- Exploring new technologies
- Integrating them into DISCOVER information system
- Hardening selected tools and making them available to the wider community

DISCOVER is a follow-on activity to our Passive Microwave ESEIP and SSM/I-SSMIS Pathfinder projects and build on user interface technology available through these programs. In addition, DISCOVER contains a Technology Development component that provides an extended set of user services and visualization tools to greatly enhance the utility of the products and increase online analysis of these data. In particular, event-driven data delivery are implemented that significantly advance distributed data capabilities. Our users require these data for their research and need support of their efforts. We solicit feedback from our users via user meetings and web-based communication tools. DISCOVER's flexibility enables it to adapt to evolving user requirements and make a significant contribution to the evolution of the SEEDS distributed data and services environment.

www.discover-earth.org

[itsc](#)

DISCOVER
www.discover-earth.org

Remote Sensing Systems
www.remss.com



Relevance of DISCOVER Technology

- **Original REASoN solicitation**

- Apply principles from the Strategy for the Evolution of ESE Data Systems (SEEDS) regarding **standards and interfaces for interoperability** and exchange of data and information
- Develop applicable **advanced data systems technologies** integrated into the project, including
 - Data and service locator technologies **leveraging the web and commercial approaches** that are tailored to the unique demands of the geo-spatial Earth science data sets
 - Techniques addressing **seamless, automated access** to data residing in **multiple distributed archives**
 - **Languages and metadata techniques** enabling Earth science community-centric services and open tool sets

- **NASA's Vision for Evolution of EOSDIS Elements**

- Seamlessly combining **multiple data and metadata streams**
- **Distributed repositories** – physical location of data irrelevant
- Service oriented architecture with **modular components** and **machine-to-machine interfaces**
- **Custom processing** to provide only the data needed, the way it is needed
- **Open interfaces** and **standard protocols** for interoperability with other relevant systems

DISCOVER Information Technology Objectives

Develop an open, distributed, heterogeneous data system designed to:

- Improve access and use of data and information products
- Extend online data distribution
- Provide for automated data management and distribution
- Provide automated order tracking and metrics

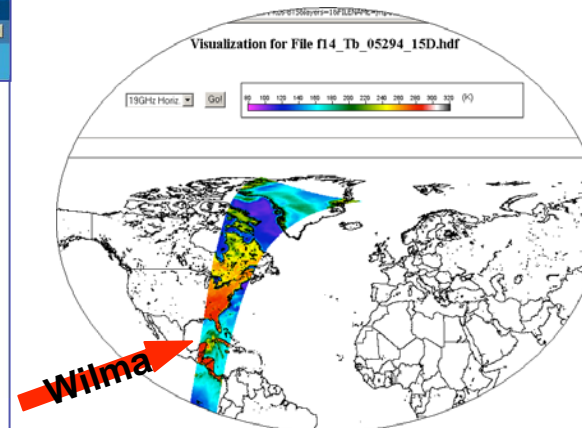
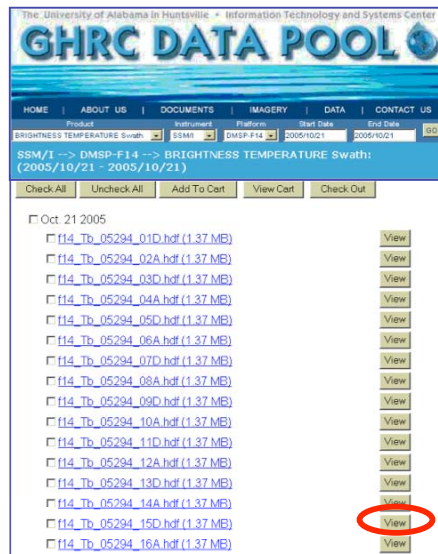
DISCOVER Information Technology Objectives

Develop a distributed, service oriented architecture for managing multiple distributed data repositories with modular components:

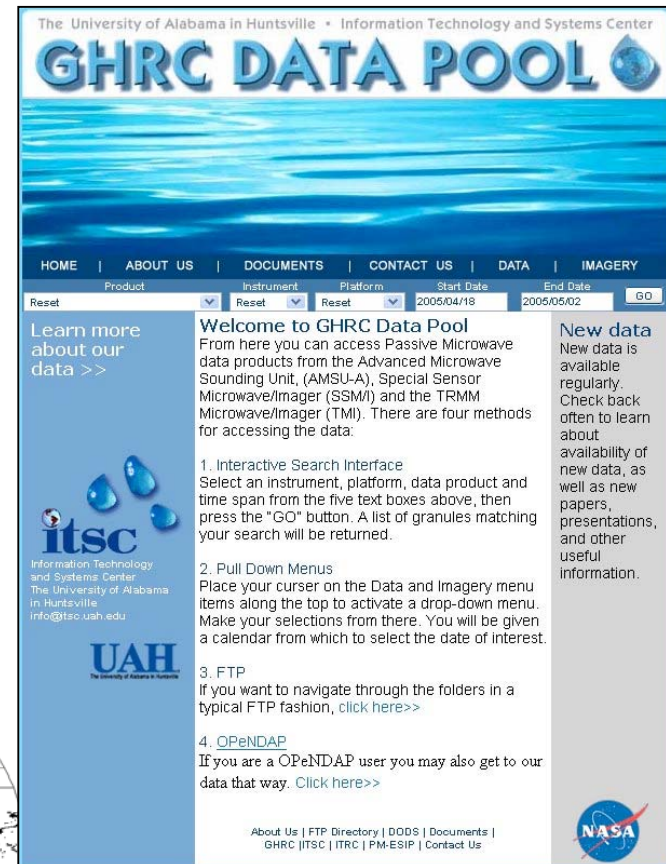
- Central Metadata Catalog Services
- Search and Order Interfaces
- Order Tracking Services
- Subsetting
- Visualization Through Web Mapping Services
- Order Broker
- Packaging Services

GHRC Data Pool

The GHRC Data Pool provides a data search-and-order interface for data in distributed repositories maintained by a single management infrastructure. The next generation of this data pool is integrating data from repositories maintained by different institutions so it has been necessary to develop tools and functionality capable of **consistently collecting and managing metadata across the distributed online resources.**



DISCOVER
www.discover-earth.org



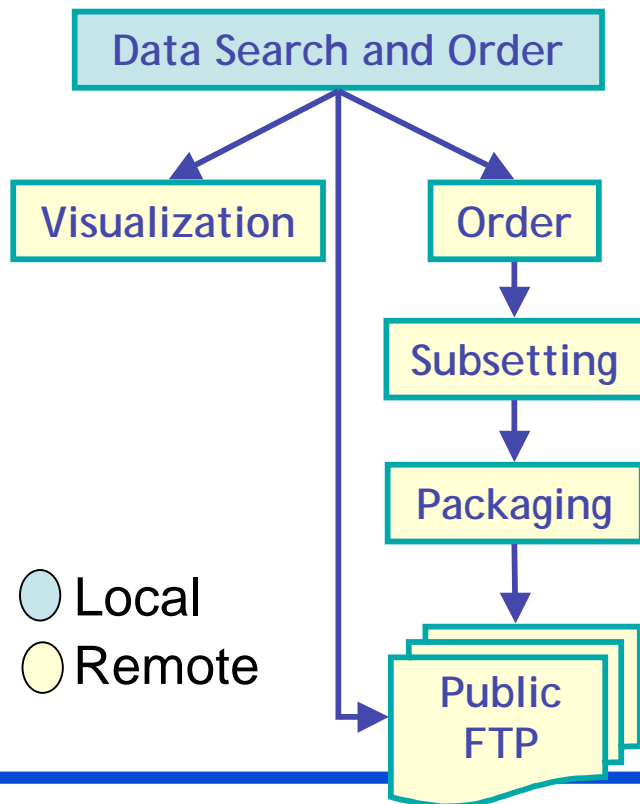
*Interactive browse images via
OGC web mapping services*

Remote Sensing Systems
www.remss.com



The University of Alabama in Huntsville • Information Technology and Systems Center

GHRC DATA POOL



- On-line data access with integrated data services
- Automated ordering, subsetting, packaging, display and delivery of scientific data
- Multiple distributed repositories at UAH and RSS
- Common user interface, **data catalog** and order tracking

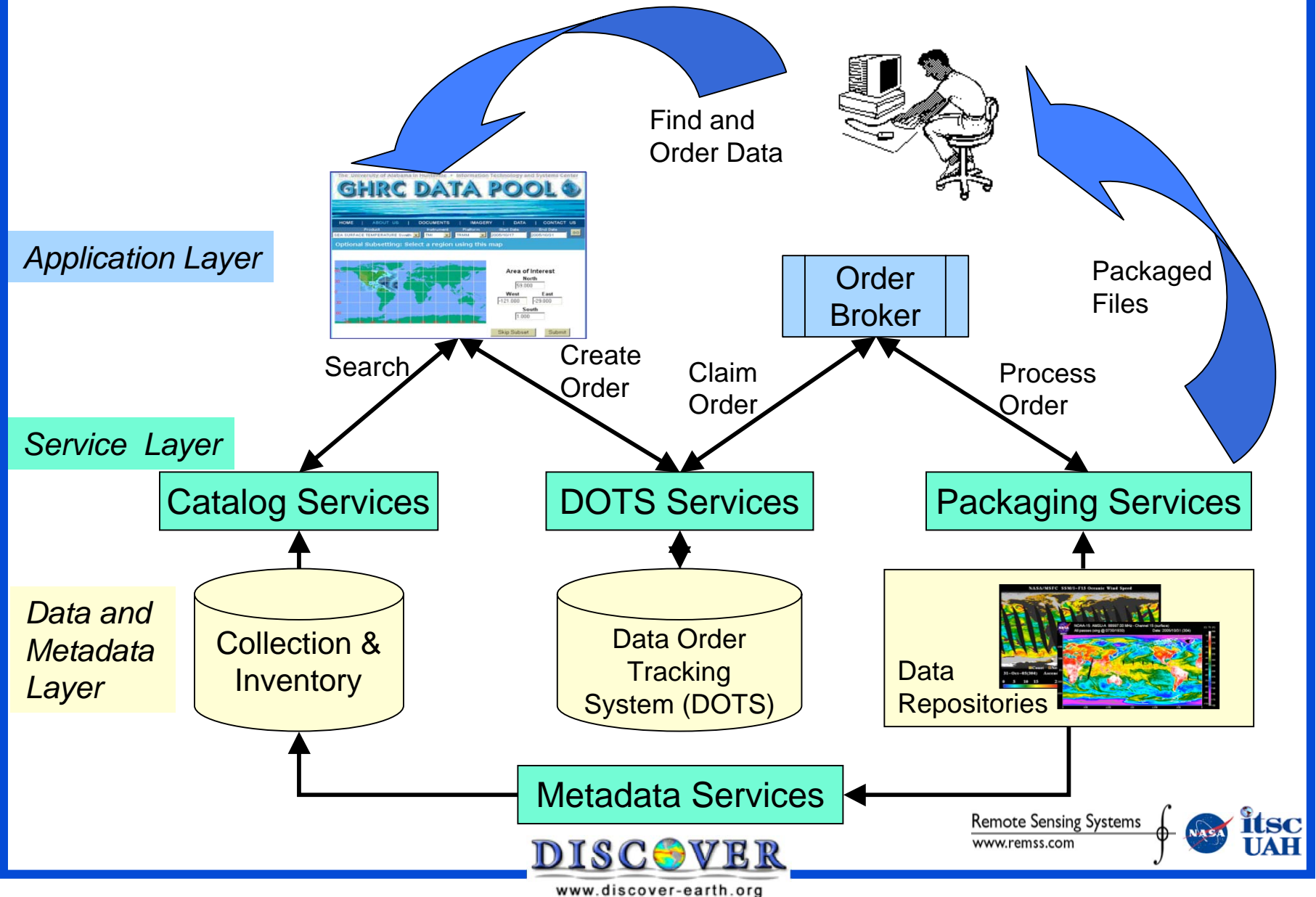
<http://datapool.nsstc.nasa.gov>

DISCOVER
www.discover-earth.org

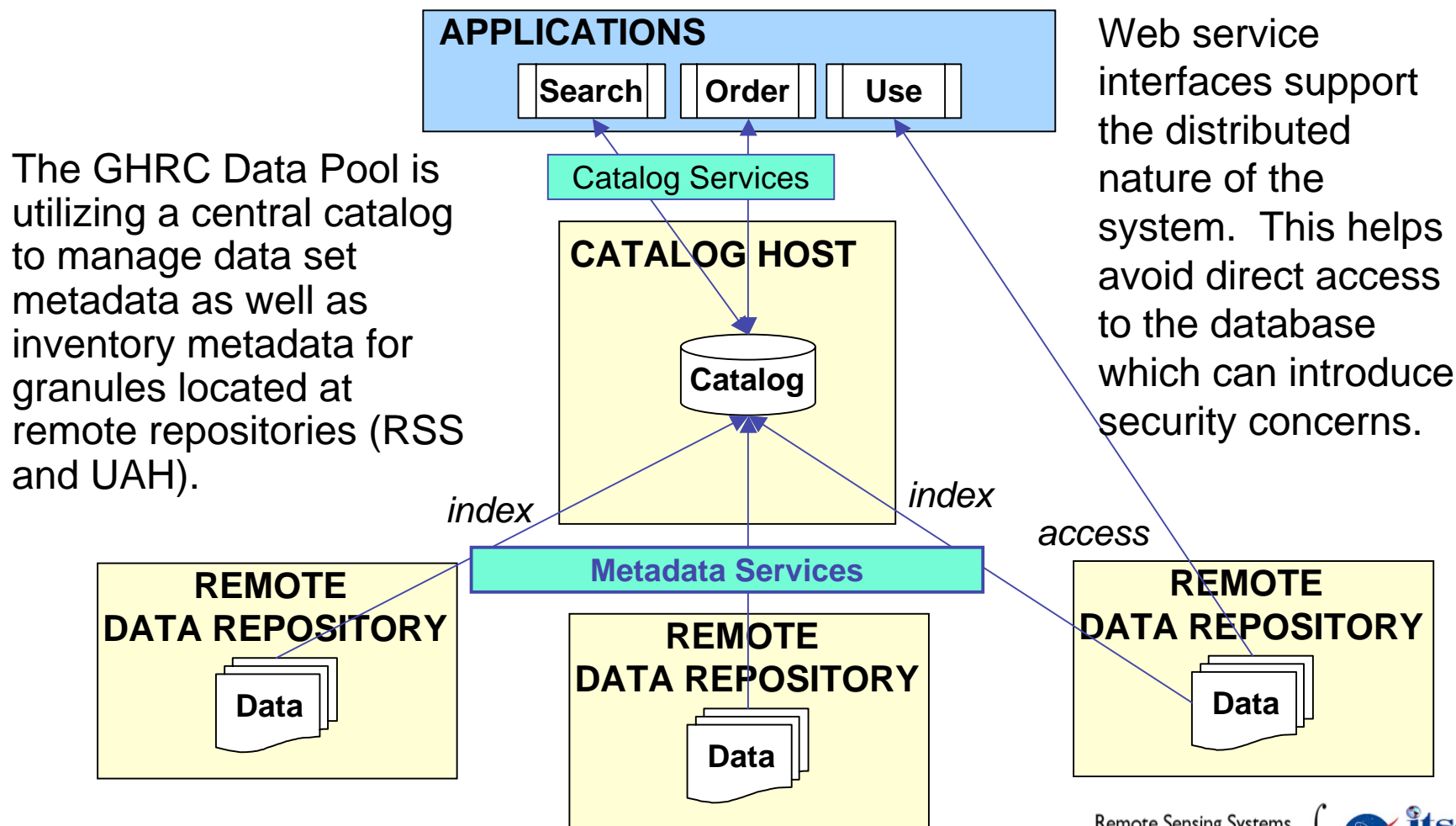
Remote Sensing Systems
www.rss.com



Data Pool Services Interactions

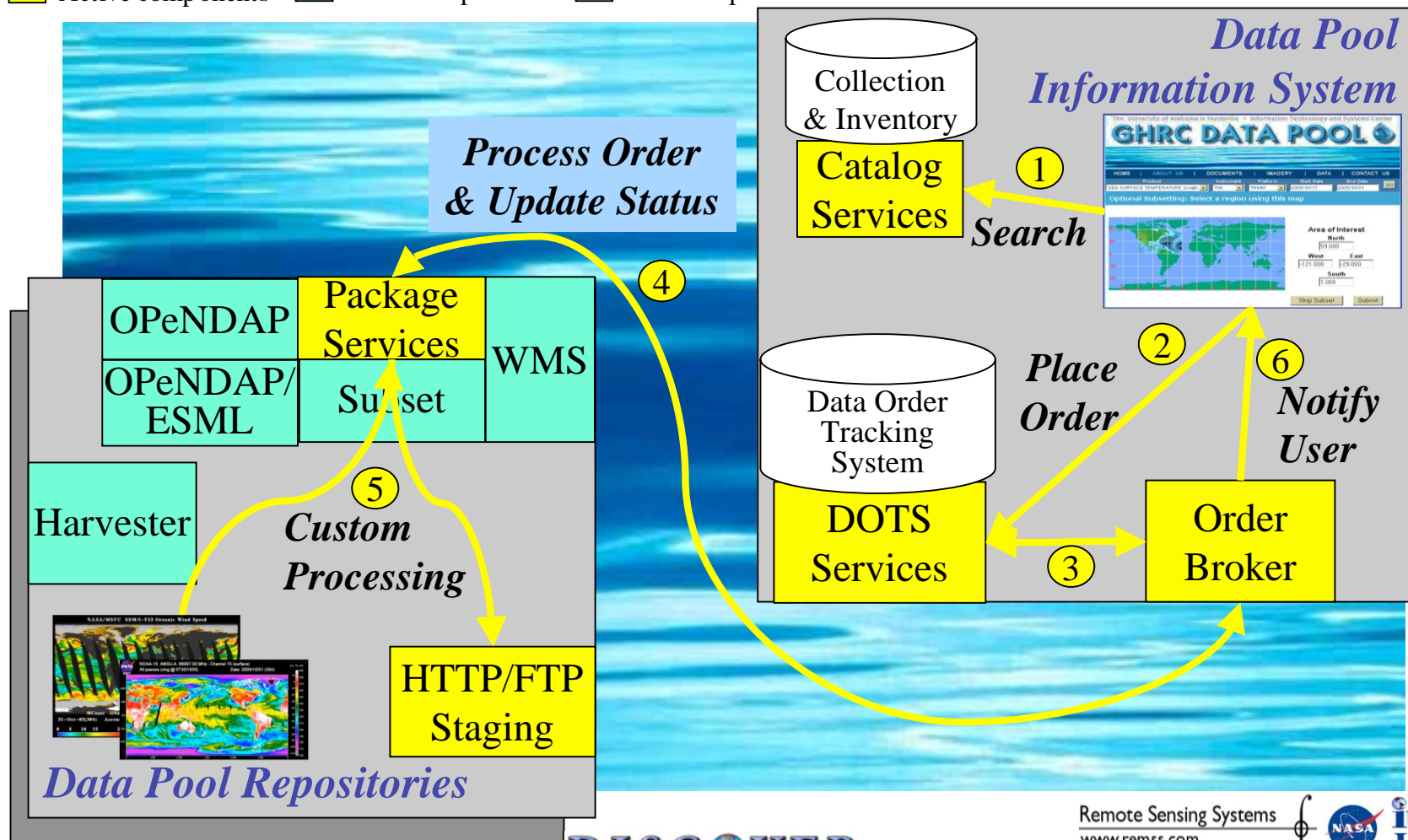


Data Pool Approach: Distributed Data Repositories



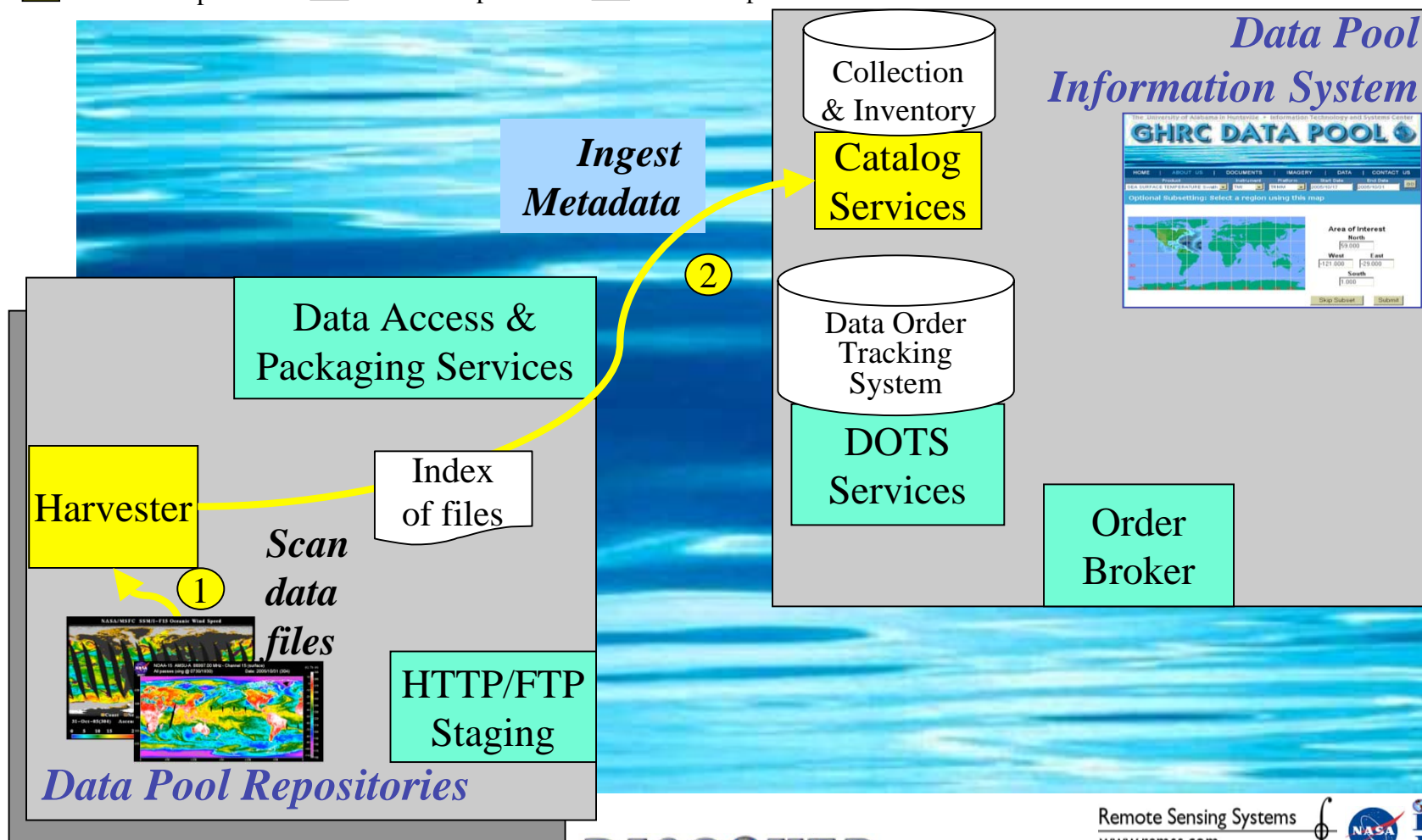
Data Pool Services Use Case: Automated Order Processing

Active components
 Other components
 Data components



Data Pool Services Use Case: Automated Metadata Ingest

■ Active components ■ Other components □ Data components

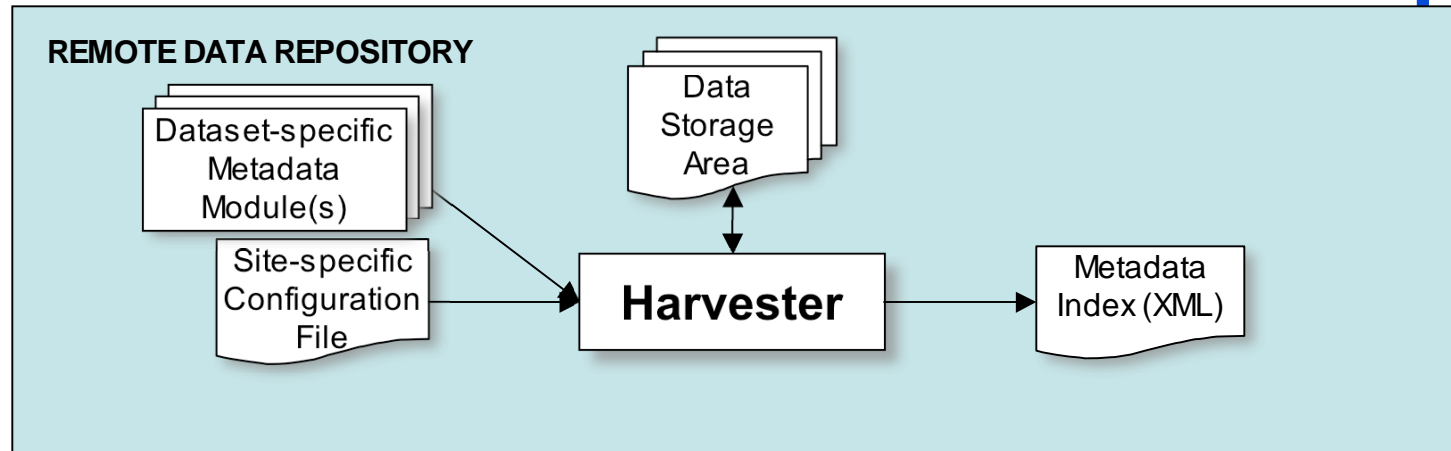


Harvester

The goal is gathering metadata for data on remote data repositories.

The Harvester is deployed on the remote repository, requiring minor site-specific configurations and dataset-specific metadata modules.

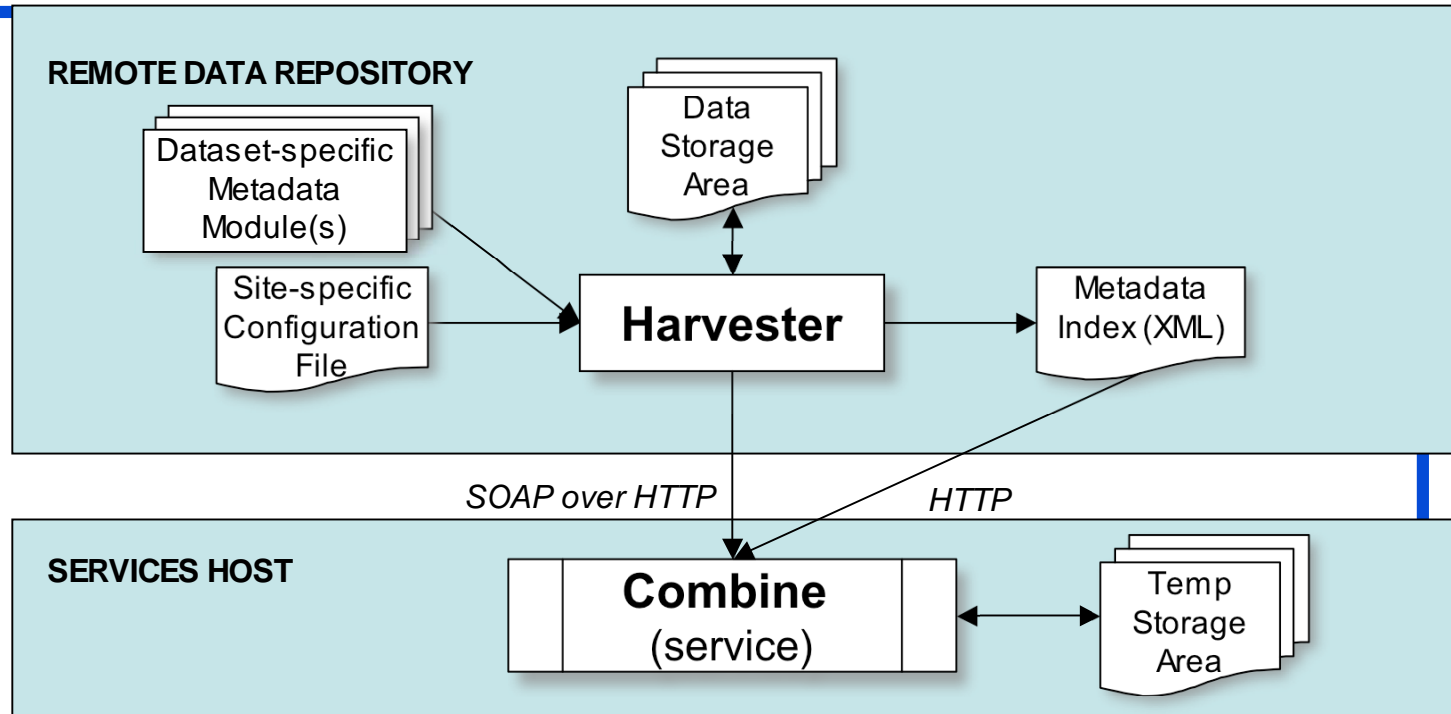
A metadata index of the remote repository is generated and maintained at the site.



Combine

The Harvester notifies a Combine at a services host (via SOAP interfaces), of the availability of the index.

The Combine retrieves the index to a local temporary work area.



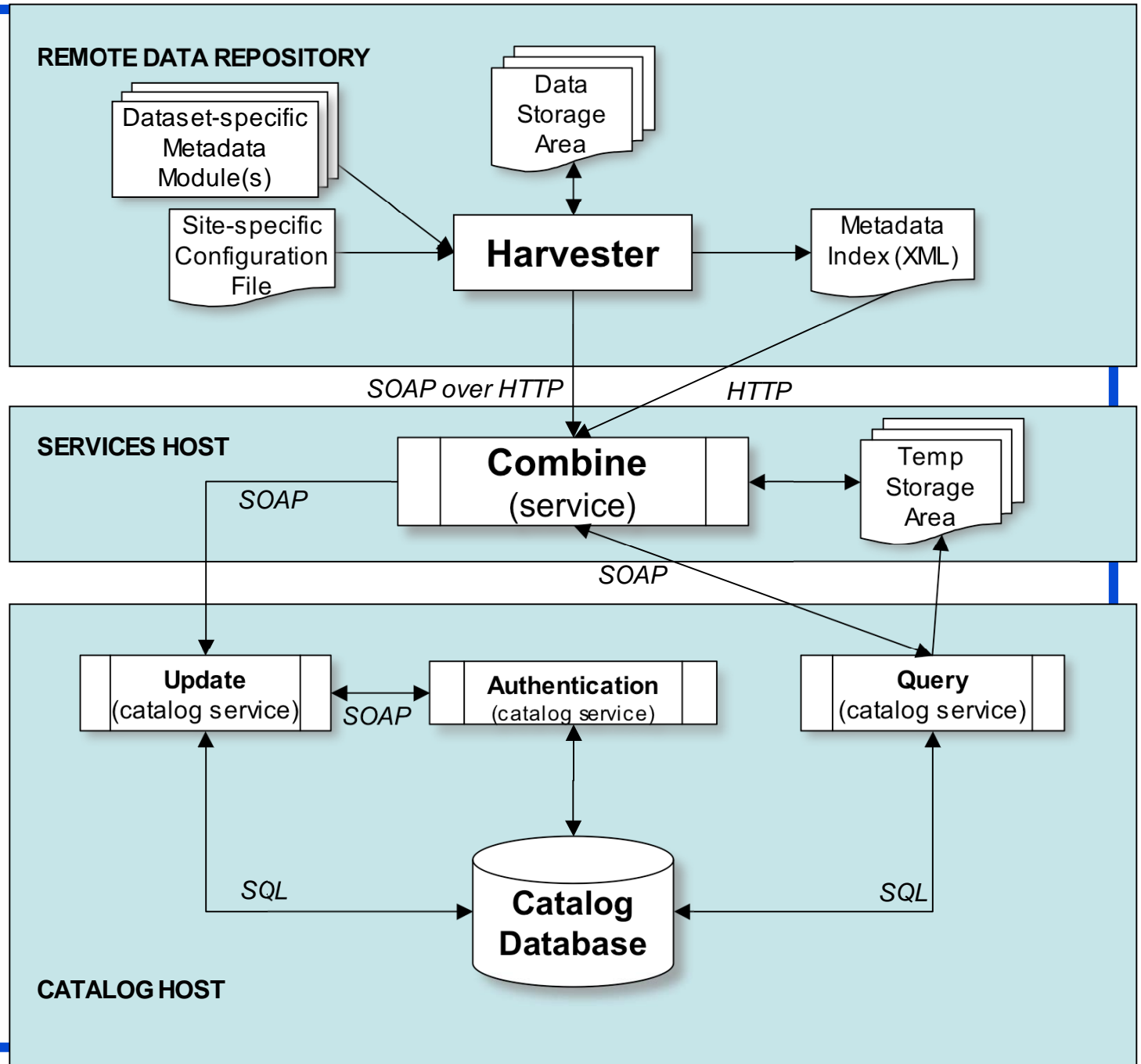
Catalog Services

Catalog services insulate the Combine from database specifics for queries and updates.

Query service returns current inventory information to Combine's temporary storage

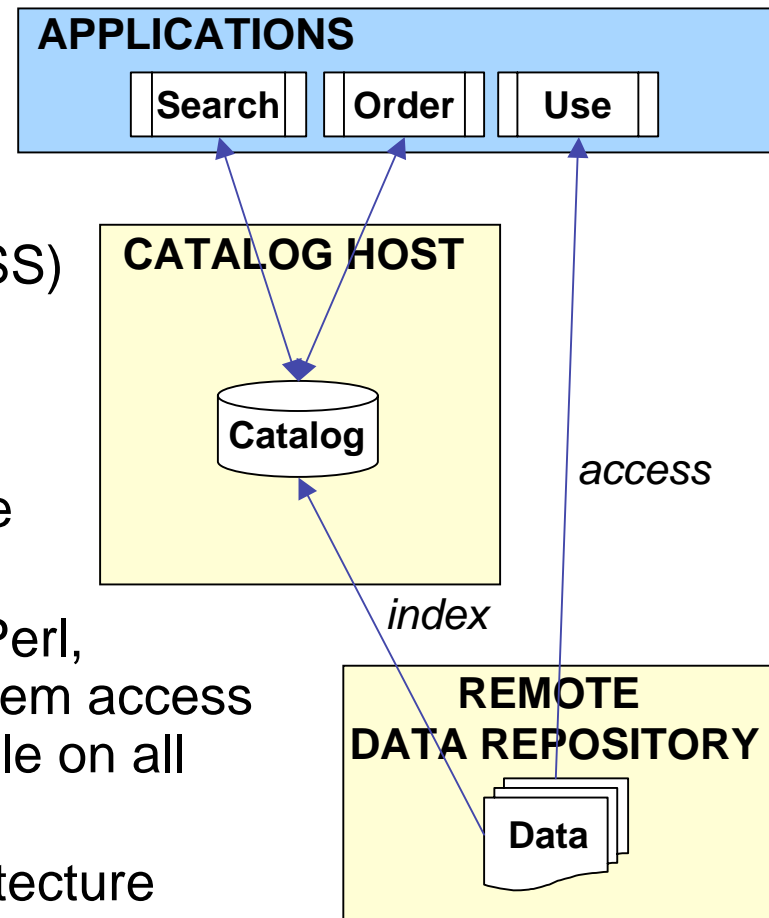
Combine cross checks the new metadata with existing inventory and determines updates needed.

Update service requires authentication before writing to the catalog database.



Conclusions

- The Harvester-Combine approach effectively utilizes a service oriented architecture to harvest inventory information for distributed online repositories (in testing at UAH and RSS)
- A services approach insulates the various distributed components from platform and language dependencies, and from changes to other parts of the system.
- The Harvester-Combine is written in Perl, utilizing platform-independent file system access modules so it has proven to be portable on all tested systems.
- This distributed service oriented architecture should prove to be a scalable solution to indexing remote repositories.



Related / Future Work

Technology research in distributed services

- Service oriented architecture for data management and processing
- Data / service semantics and knowledge management
- Techniques to transfer large binary data between services
- Service discovery and workflow formulation
- Automated service invocation
- Express data delivery services via subscription

Acknowledgements and Links

DISCOVER partners at the National Space Science and Technology Center (NSSTC) and at Remote Sensing Systems have assisted in the deployment and testing of these components.

DISCOVER Project: <http://www.discover-earth.org>

DISCOVER IT: <http://discover.itsc.uah.edu>

GHRC Data Pool: <http://datapool.nsstc.nasa.gov>